# Co-clustering contraint pour le résumé de matrices document-terme

Margot Selosse [1] & Julien Jacques [2] & Christophe Biernacki [3]

[1] *5 Avenue Pierre Mendès France, 69500 Bron margot.selosse@gmail.com*
[2] *5 Avenue Pierre Mendès France, 69500 Bron julien.jacques@univ-lyon2.fr*
[3] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq Lille christophe.biernacki@inria.fr*

**Résumé.** Le co-clustering est une méthode de fouille de données qui produit simultanément un clustering des observations (en ligne) et un clustering des variables (en colonne). Ce travail présente un nouveau modèle de co-clustering pour résumer des données textuelles stockées sous la forme de matrice document-terme. Nous appelons bloc le croisement d'un cluster en ligne et d'un cluster en colonne. Notre modèle met en évidence des blocs homogènes, mais distingue aussi les blocs significatifs des blocs dits "de bruit". Cela est particulièrement utile pour les matrices document-terme qui sont sparses et de haute dimension. De plus, le modèle propose une organisation parmi les blocs significatifs et de bruit, ce qui rend plus facile pour l'utilisateur d'interpréter les résultats. Un algorithme Stochastic Gibbs Expectation-Maximization (SEM-Gibbs) est utilisé pour l'inférence du modèle.

**Mots-clés.** Modèle des blocs latents, données textuelles, interprétabilité

**Abstract.** Co-clustering is a data mining technique which simultaneously produces row-clusters of observations and column-clusters of features. This work presents a novel co-clustering model which easily summarizes textual data in a document-term format. In addition to highlighting homogeneous co-clusters, we also distinguish noisy co-clusters from significant co-clusters, which are particularly useful for sparse document-term matrices. Furthermore, the model proposes a structure among the significant co-clusters, thus providing improved interpretability to users. A Stochastic Expectation-Maximization algorithm is proposed to implement the model's inference as well as a model selection criterion to choose the number of co-clusters.

**Keywords.** Latent Block Model, textual data, interpretability

# 1 Introduction

Ce travail présente le modèle SOCC (Self Organised Co-Clustering). Son objectif est de résumer des matrices document-terme, dont les lignes correspondent à des documents, et dont les colonnes correspondent à des mots (ou termes). Une cellule $(i, j)$ correspond

au nombre de fois que le $j$-ème terme apparaît dans le $i$-ème document. Le clustering, qui forme des groupes homogènes d'observations (de documents dans notre cas), est une technique non-supervisée qui a prouvé son efficacité dans plusieurs domaines. Cependant, dans des contextes sparses et de haute dimension, les techniques de clustering classiques sont moins adaptées et difficiles à interpréter. Avec de tels jeux de données, le co-clustering - qui regroupe les observations et les variables simultanément - peut se révéler plus efficace. Le co-clustering permet de résumer les jeux de données en blocs (le croisement entre un cluster en ligne et un cluster en colonne). Dans le cadre de matrices document-terme, les clusters de documents aident à trouver les documents qui parlent du même sujet, tandis que les clusters de termes aident à savoir de quoi parlent ces documents.

Nous nous intéressons ici à une approche probabiliste appelée le modèle des blocs latents [3]. Elle suppose que les données sont générées depuis un mélange de distributions de probabilités, dont chaque composant correspond à un bloc. Les paramètres des distributions correspondantes et les appartenances aux blocs sont ensuite estimés à partir des données. Cette approche modélise les éléments d'un bloc avec une distribution paramétrique: chaque bloc est interprétable à partir de ses paramètres de distribution.

Toutefois, lorsqu'il s'agit de données sparses et de haute dimension, plusieurs blocs peuvent être extrêmement sparses (composés de zéros) et causer des problèmes d'inférence. En outre, la mise en évidence de blocs homogènes n'est pas toujours suffisante pour obtenir des résultats faciles à interpréter. En effet, malgré leur homogénéité, ces blocs sparses ne sont pas significatifs du point de vue de l'interprétation, et nous avons besoin d'une nouvelle étape pour différencier les blocs significatifs des autres. En d'autres termes, il est laissé à l'utilisateur de choisir les blocs les plus utiles et de déterminer quels clusters de termes (clusters en colonnes) sont plus spécifiques à quels clusters de documents (clusters en lignes). Cette tâche n'est pas triviale, même avec un nombre raisonnable de clusters en lignes et de colonnes. Il est donc nécessaire de travailler sur une technique de co-clustering qui offre des résultats prêts à l'emploi.

# 2 Co-clustering et le modèle des blocs latents

## 2.1 Notations

Nous considérons une matrice $X$ avec un nombre $I$ de lignes et de $J$ colonnes. Nous notons $x_{ij}$ un élément de $X$ tel que $1 \leq i \leq I$ et $1 \leq j \leq J$. Etant dans un contexte de co-clustering, nous supposons qu'il existe $G$ clusters en ligne, $H$ clusters en colonnes. Pour cela, nous introduisons les matrices $\boldsymbol{v}$ et $\boldsymbol{w}$, qui correspondent respectivement aux partitions des clusters en lignes et colonnes. Ainsi $v_i$, la ième ligne de $\boldsymbol{v}$, est un vecteur de taille $G$, tel que $v_{ig}$ est égal à 1 lorsque la $i$-ème ligne appartient au $g$-ième cluster en ligne, et 0 dans le cas contraire. De la même manière, $w_j$, la j-ème ligne de $\boldsymbol{w}$ est un vecteur de taille $H$ tel que $w_{jh}$ est égal à 1 lorsque la $j$-ème colonne appartient $h$-ième cluster en colonne, et 0 autrement.

## 2.2 Hypohèses du modèle

Le modèle des blocs latents [3] se base sur les hypothèses suivantes:

**Hypothèse 1** *Les partitions $v_i, w_j$ sont indépendantes pour tout $\{i, j\}$.*

Cela se traduit donc par :

$$p(\boldsymbol{v}, \boldsymbol{w}) = p(\boldsymbol{v})p(\boldsymbol{w}) = \prod_i p(v_i) \prod_j p(w_j) = \prod_{ig} \alpha_g^{v_{ig}} \prod_{jh} \beta_h^{w_{jh}}, \tag{1}$$

où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_G)$ et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_H)$ sont les proportions de mélange des clusters, en ligne et en colonne respectivement.

**Hypothèse 2** *Conditionnellement à $\boldsymbol{v}$ et $\boldsymbol{w}$, les éléments $x_{ij}$ d'un bloc sont indépendants et identiquement distribués.*

Nous avons donc :

$$x_{ij}|v_{ig}w_{jh} = 1 \overset{iid}{\sim} f(.; \theta_{gh}) \text{ pour tout } (i, j).$$

Ici, $\theta_{gh}$ représente le paramètre de la distribution du bloc formé par les clusters en lignes $g$ et les clusters en colonnes $h$. Par la suite, nous l'appellerons simplement le bloc $(g, h)$.

Ainsi, nous obtenons :

$$p(X|\boldsymbol{v}, \boldsymbol{w}) = \prod_{ijgh} f(x_{ij}; \theta_{gh})^{v_{ig}w_{jh}}.$$

La vraisemblance du modèle peut donc être écrite :

$$p(X) = \sum_{\substack{(\boldsymbol{v},\boldsymbol{w}) \\ \in V \times W}} p(X|\boldsymbol{v}, \boldsymbol{w})p(\boldsymbol{v}, \boldsymbol{w}) = \sum_{\substack{(\boldsymbol{v},\boldsymbol{w}) \\ \in V \times W}} \prod_{ig} \alpha_g^{v_{ig}} \prod_{jh} \beta_g^{w_{jh}} \prod_{ijgh} f(x_{ij}; \theta_{gh})^{v_{ig}w_{jh}}, \tag{2}$$

où $V$ et $W$ sont l'ensemble des partitions possibles.

## 2.3 Le co-clustering avec la distribution de Poisson

Comme $x_{ij}$ dénombre le nombre d'occurence du mot $j$ dans le document $i$, la modélisation par une loi de Poisson est naturelle. Dans ce contexte, il est considéré qu'un élément $x_{ij}$ est tiré d'une loi de Poisson de paramètre $\lambda_{ij}$, soit:

$$x_{ij} \sim \mathcal{P}(\lambda_{ij}).$$

Ainsi, nous avons:

$$f(x_{ij}; \lambda_{ij}) = e^{-\lambda_{ij}} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}.$$

Le paramètre $\lambda_{ij}$, lui, est considéré être une fonction d'un effet de bloc $\theta_{gh}$, d'un effet de ligne $\mu_i$ et d'un effet de colonne $\nu_j$:

$$\lambda_{ij} = \mu_i \nu_j \sum_{gh} v_{ig} w_{jh} \theta_{gh}.$$

Pour assurer l'identifiabilité du modèle, nous fixons $\mu_i$ et $\nu_j$ tel que suit:

$$\mu_i = \sum_j x_{ij}, \text{ et } \nu_j = \sum_i x_{ij}.$$

En fixant $\mu_i$ et $\nu_j$, le seul paramètre à estimer concernant la distribution de Poisson est $\boldsymbol{\theta} = (\theta_{gh})$.

# 3 Le modèle Self-organised Co-Clustering (SOCC)

## 3.1 Le co-clustering contraint SOCC

Jusqu'à maintenant, nous avons décrit un modèle des blocs latents classique, avec utilisation d'une distribution de Poisson. Les paramètres $\theta_{gh}$ sont sans rapport et donc chaque bloc doit être interprété séparément des autres. Dans le modèle SOCC, cette indépendance n'est plus supposée. Ainsi, pour un bloc donné $(g, h)$, l'effet de bloc correspondant $\theta_{gh}$ sera soit spécifique au cluster en colonnes $h$ avec $\theta_{gh} = \theta_h$, soit non-spécifique, avec $\theta_{gh} = \theta$. Dans le cas d'un effet de bloc non-spécifique $\theta_{gh} = \theta$, le bloc $(g, h)$ est considéré comme un bloc de bruit, et il partage le même paramètre $\theta$ avec tous les autres blocs de bruit. Dans le cas de $\theta_{gh} = \theta_h$, le bloc $(g, h)$ est significatif, et partage le même $\theta_h$ avec tous les blocs significatifs du même cluster en colonnes $h$. Dans ce cas, les termes du $h$-ième cluster en colonne sont considérés comme spécifiques aux documents d'un ou plusieurs clusters en lignes.

Pour organiser les blocs significatifs et les blocs de bruit, plusieurs règles sont données. Tout d'abord, après avoir choisi le nombre de clusters en lignes $G$, le nombre de clusters en colonnes est égal à $H = G + \binom{G}{2} + 1$. De plus, les clusters en colonnes sont divisés en trois parties appelées *main*, *second* et *common*. Ces parties sont illustrées par la Figure 1 et leur objectif est expliqué ici. La partie *main* concerne les $G$ premiers clusters en colonnes, pour $h \in \{1, ..., G\}$. Dans chaque cluster en colonnes $h$ de cette partie, un seul bloc est significatif et paramétré par $\theta_h$. Tous les autres blocs sont de bruit et paramétrés par $\theta$. Par conséquent, pour chaque cluster de documents, le bloc significatif indique les termes qui sont spécifiques à ces documents. Ainsi dans la partie *main* les blocs significatifs sont en diagonale, et les autres blocs sont des blocs de bruit. La partie *second* concerne les $\binom{G}{2}$ clusters en colonnes suivants ($h \in \{G + 1, ..., G + \binom{G}{2}\}$). Dans chaque cluster en colonnes $h$ de cette partie, deux blocs sont considérés significatifs. Chaque cluster en colonnes contient donc des termes spécifiques à deux clusters de documents. Enfin,

la partie *common* ne comprend qu'un seul cluster en colonnes et rassemble les termes communs à tous les documents.

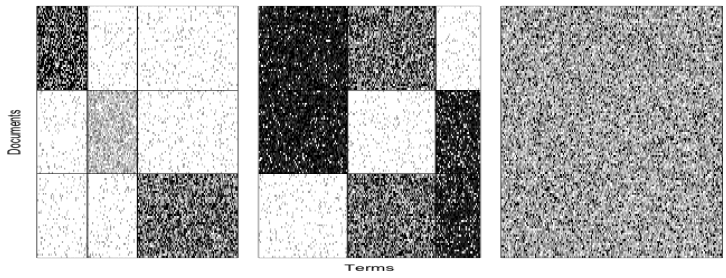


Figure 1: Illustration du co-clustering effectué avec le modèle SOCC. De gauche à droite, la partie *main*, la partie *second* et la partie *common*.

**Illustration** Dans la Figure 1, nous voyons clairement les blocs significatifs (les plus foncés) avec $\theta_{gh} = \theta_h$ et les blocs de bruit (les plus clairs) avec $\theta_{gh} = \theta$. Nous discernons également l'organisation entre ces blocs et les trois différentes parties *main*, *second* et *common*. Par exemple, dans la partie *main*, le premier cluster en colonnes est considéré comme spécifique au premier cluster en ligne, ainsi seul le premier bloc du cluster en colonnes a sa propre distribution spécifique avec $\theta_1$. En revanche, les autres blocs de ce cluster en colonnes sont considérés comme de bruit et ont un paramètre d'effet de bloc $\theta$, qui est commun à tous les blocs de bruit. Dans la partie *second*, nous observons que pour $h = 4$, les blocs $(1, 4)$ et $(2, 4)$ sont significatifs, et partagent le même effet de bloc $\theta_4$. Cela implique que les termes de cluster en colonne 4 sont spécifiques aux documents des clusters en ligne 1 et 2. De plus, le bloc $(4, 3)$ est de bruit et a le même effet $\theta$ que les autres blocs de bruit. La partie *common* est particulière dans la mesure où elle ne contient qu'un seul cluster en colonnes, donc $h = 7$. Ce cluster en colonnes contient les termes spécifiques à tous les clusters de documents et ses blocs correspondants partagent tous le même $\theta_7$.

En notant $\mathcal{C}_h$ les blocs significatifs du cluster en colonnes $h$ et $\overline{\mathcal{C}}_h$ les blocs de bruit du cluster en colonnes $h$, la probabilité du modèle SOCC s'écrit :

$$p(X) = \sum_{\substack{(\boldsymbol{v},\boldsymbol{w}) \\ \in V \times W}} \prod_{ig} \alpha_g^{v_{ig}} \prod_{jh} \beta_h^{w_{jh}} \prod_{ijh} \prod_{g \in \mathcal{C}_h} f(x_{ij}; \theta_h)^{v_{ig} w_{jh}} \prod_{g \in \overline{\mathcal{C}}_h} f(x_{ij}; \theta)^{v_{ig} w_{jh}}. \tag{3}$$

## 3.2 Inférence du modèle et choix de $G$

**Utilisation d'une alternative à l'algorithme EM (Expectation-Maximisation).** Pour estimer les variables latentes $\boldsymbol{v}$ et $\boldsymbol{w}$ et les paramètres $\boldsymbol{\theta}$, l'algorithme Expectation-

Maximisation (EM) [2] semble être un bon candidat. Cependant, dans le cadre du co-clustering, cet algorithme nécessite de calculer $p(v_{ig}w_{jh} = 1|X)$. Or, cette quantité n'est pas facilement calculable, ce qui rend l'utilisation d'un EM classique impossible. Nous utilisons alors un algorithme appelé Stochastic-Gibbs Expectation-Maximisation (SEM-Gibbs) [4].

**Utilisation d'un critère ICL pour sélectionner le nombre de cluster.** Comme le nombre de clusters en colonnes $H$ est directement induit du nombre de clusters en ligne $G$ (on a $H = G + \binom{G}{2} + 1$), $G$ est le seul nombre à choisir pour utiliser le modèle SOCC. En co-clustering, le critère de sélection BIC (Bayesian Information Criterion) [5] n'est pas calculable non plus. Pour connaître le nombre $G$ optimal, nous utilisons le critère de sélection ICL (Integrated Complete Likelihood) [1].

# 4 Conclusion

Le modèle SOCC est un nouvel algorithme de co-clustering, adapté pour les matrices document-terme, et basé sur le modèle des blocs latents. Nous comparerons les résultats du modèle avec d'autres techniques sur des jeux de données réels. Nous montrerons aussi les résultats obtenus avec pour données d'étude les trois premiers tomes de Harry Potter.

# References


[1] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 22(7):719–725, July 2000.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of he Royal Statistical Society, series B*, 39(1):1–38, 1977.

[3] G. Govaert and M. Nadif. Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39(3):416–425, 2010.

[4] Christine Keribin, Gérard Govaert, and Gilles Celeux. Estimation d'un modèle à blocs latents par l'algorithme SEM. In *42èmes Journées de Statistique*, Marseille, France, France, 2010.

[5] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.